Pipelines de ETL e Machine Learning com Apache Spark

Componentes do Cluster Spark Master, Worker e History Server

O Apache Spark possui uma arquitetura modular que inclui vários componentes. Vamos nos concentrar nesses três: Master, Worker e History Server.

Master (ou Spark Master):

- Função: O Master é o ponto central do cluster Spark que coordena a distribuição de aplicações (jobs) no cluster.
- Responsabilidades: Ele aloca recursos (como memória e CPU) para cada aplicação com base na configuração fornecida e nas demandas do cluster. Também rastreia a disponibilidade dos Workers e os recursos que eles têm.
- UI (User Interface): O Master também fornece uma interface web para visualização do status e gerenciamento do cluster.
- Endereço padrão: Por padrão, a UI do Spark Master pode ser acessada no endereço http://[hostname]:8080.

Worker (ou Spark Worker):

- Função: Os Workers são os nós do cluster onde o trabalho é realizado.
- Responsabilidades: Cada Worker é responsável por executar as tarefas atribuídas a ele pelo Master. Os Workers também informam regularmente o Master sobre os recursos disponíveis e o status das tarefas.
- Executor: Dentro de cada Worker, os executores são inicializados para cada aplicação. Esses executores são processos JVM (Java Virtual Machine) separados que executam tarefas para uma aplicação Spark específica.
- Armazenamento: Os Workers também podem armazenar dados em cache ou intermediários em sua memória ou disco local, conforme instruído pela aplicação.

History Server:

- Função: O Spark History Server fornece uma interface web para visualizar os logs e métricas de aplicações Spark concluídas.
- Responsabilidades: Depois que uma aplicação Spark é concluída, suas métricas e logs podem ser armazenados e analisados posteriormente através do History Server. Isso é útil para análises pós-execução, otimização e depuração.
- Armazenamento: Os logs das aplicações (também conhecidos como logs de eventos) são normalmente armazenados em um sistema de arquivos distribuído como o HDFS ou em um sistema de arquivos local, de onde o History Server pode lê-los.
- Endereço por padrão: A UI do History Server geralmente pode ser acessada no endereço http://[hostname]:18080.

Resumindo: O Master coordena o cluster, os Workers realizam o trabalho real e o History Server permite a revisão e análise de aplicações concluídas. Esses componentes juntos permitem que o Spark gerencie e execute tarefas em um ambiente distribuído de maneira eficiente.

Vamos agora construir um Cluster Spark com esses 3 componentes.

Equipe DSA

Muito Obrigado!
Continue Trilhando Uma Excelente Jornada de Aprendizagem.